Ultradeep Bisulfite Sequencing Analysis of DNA Methylation Patterns in Multiple Gene Promoters by 454 Sequencing

Kristen H. Taylor,[1] Robin S. Kramer,[2] J. Wade Davis,[3] Juyuan Guo,[1] Deiter J. Duff,[1] Dong Xu,[2] Charles W. Caldwell,[1] and Huidong Shi[1]

[1]Department of Pathology and Anatomical Sciences, Ellis Fischel Cancer Center; [2]Department of Computer Sciences and Christopher S. Bond Life Sciences Center; and [3]Department of Statistics and Health Management and Informatics, University of Missouri-Columbia, Columbia, Missouri

Abstract

We developed a novel approach for conducting multisample, multigene, ultradeep bisulfite sequencing analysis of DNA methylation patterns in clinical samples. A massively parallel sequencing-by-synthesis method (454 sequencing) was used to directly sequence >100 bisulfite PCR products in a single sequencing run without subcloning. We showed the utility, robustness, and superiority of this approach by analyzing methylation in 25 gene-related CpG rich regions from >40 cases of primary cells, including normal peripheral blood lymphocytes, acute lymphoblastic leukemia (ALL), chronic lymphocytic leukemia (CLL), follicular lymphoma (FL), and mantle cell lymphoma (MCL). A total of 294,631 sequences was generated with an average read length of 131 bp. On average, >1,600 individual sequences were generated for each PCR amplicon far beyond the few clones (<20) typically analyzed by traditional bisulfite sequencing. Comprehensive analysis of CpG methylation patterns at a single DNA molecule level using clustering algorithms revealed differential methylation patterns between diseases. A significant increase in methylation was detected in ALL and FL samples compared with CLL and MCL. Furthermore, a progressive spreading of methylation was detected from the periphery toward the center of select CpG islands in the ALL and FL samples. The ultradeep sequencing also allowed simultaneous analysis of genetic and epigenetic data and revealed an association between a single nucleotide polymorphism and the methylation present in the *LRP1B* promoter. This new generation of methylome sequencing will provide digital profiles of aberrant DNA methylation for individual human cancers and offers a robust method for the epigenetic classification of tumor subtypes. [Cancer Res 2007;67(18):8511–8]

Introduction

Epigenetic processes control the packaging and function of the human genome and contribute to normal and pathologic states, including cancer. It is increasingly being recognized that epigenetic alterations play a major role in driving tumor initiation and progression. The main processes that contribute to the epigenome of a cell are DNA methylation and histone modifications. Methylation of cytosine residues at CpG dinucleotides is the best-studied epigenetic modification in mammalian genomes and is known to have profound effects on gene expression. This epigenetic event occurs globally in the normal genome, affecting 70% to 80% of all CpG dinucleotides (1, 2). However, ~1-kb stretches of GC-rich DNA called CpG islands (CGI), most of which are located within or near regulatory regions of genes, seem to be protected from this modification in normal somatic cells (3). Accumulating evidence indicates that CGI methylation within promoters of specific tumor suppressor genes is associated with transcriptional silencing that compromises control of cell proliferation (4). Thus, aberrant methylation is now being investigated as a potential biomarker and for the development of pathway-specific therapeutic targets.

Bisulfite genomic sequencing, which examines multiple subclones of a bisulfite PCR product, is time consuming and is potentially affected by bias (5) and heteroduplex amplification (6). In this method, only a few clones are typically analyzed (<20), which results in a SE of the estimate of methylation that is generally excessively wide. Furthermore, because the clones are typically obtained from only a few biological samples, it is difficult to generalize the results to a larger population. To reduce the bias introduced by subcloning, the human epigenome project in Europe used direct sequencing of bisulfite PCR products (7, 8). In this method, the methylation present at any given CpG site is estimated by taking the average of all fragments (thousands) generated during PCR and results in a more statistically robust representation of the methylation present compared with subcloning. However, this approach is not without limitations. For example, it is not possible to determine the methylation patterns of individual DNA molecules, and there is a lack of sensitivity in cases involving low levels of methylation.

Recently, a novel massively parallel sequencing-by-synthesis method was commercialized that is based on pyrosequencing in picoliter-scale reactions (454 sequencing). Approximately 300,000 DNA templates can be simultaneously sequenced in a single 5.5-h run to an average read length of 100 bases, with an accuracy of 99.6% (9). This highly parallel sequencing system potentially has many important applications. In this study, we investigated the use of this technology for bisulfite genomic sequencing. Due to the vast throughput capabilities of this technology, we were able to conduct a multisample, multigene, ultradeep bisulfite sequencing analysis in primary lymphoma and leukemia samples. The results indicate that this large-scale genomic bisulfite sequencing approach will provide an efficient method for deeply exploring the human cancer epigenome.

Materials and Methods

DNA pooling and bisulfite treatment. Tissue and blood samples of patients diagnosed with precursor B-cell acute lymphoblastic leukemia

---

Note: Supplementary data for this article are available at Cancer Research Online (http://cancerres.aacrjournals.org/).

K.H. Taylor and R.S. Kramer contributed equally to this work.

Requests for reprints: Huidong Shi or Charles W. Caldwell, Department of Pathology and Anatomical Sciences, University of Missouri-Columbia, Columbia, MO 65212. Phone: 573-882-5523; E-mail: shihu@health.missouri.edu or caldwellc@health.missouri.edu.

doi:10.1158/0008-5472.CAN-07-1016

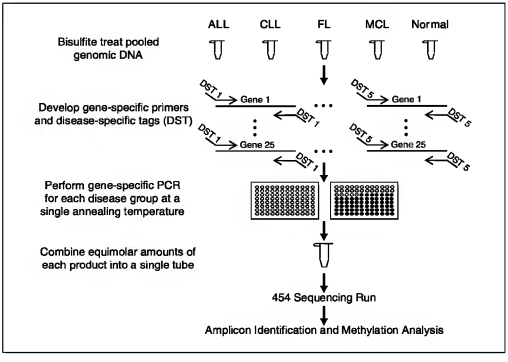

Figure 1. Experimental design for bisulfite sequencing using 454 Genome Sequencer 20.

(ALL) or non–Hodgkin's lymphomas at the Ellis Fischel Cancer Center (Columbia, MO) were obtained in compliance with the local Institutional Review Board. DNA was isolated using the QIAamp DNA Mini kit (Qiagen) from a total of 40 specimens: 10 of each from mantle cell lymphoma (MCL), chronic lymphocytic leukemia (CLL), follicular lymphoma (FL), and ALL (Supplementary Table S1). Equal amounts of DNA from each patient sample were pooled by disease type. Commercial peripheral blood lymphocyte (PBL) DNAs derived from a pool of multiple individuals (two different female pools and two different male pools) were combined and used to create the DNA pool referred to as normal PBL. DNA (1 μg) from each pool was treated with sodium bisulfite using an EpiTect kit (Qiagen).

Amplicon design and PCR. Disease-specific primers were developed for 25 genes using MethPrimer. Each gene studied has a core primer sequence that is identical across all groups. In addition, a four-nucleotide disease-specific tag was added to the 5′ end of each core primer sequence so that each group could be computationally separated after 454 sequencing analysis. Each primer pair was amplified in a 25 μL reaction using a touchdown PCR at annealing temperatures from 60°C to 56°C (one cycle at each temperature) followed by 30 cycles at an annealing temperature of 56°C. Denaturation (95°C), annealing, and extension (72°C) times are 15 s, 30 s, and 1 min, respectively. For each gene, five disease-specific amplicons were generated by PCR. Each amplicon was individually prepared, purified, and quantified. One hundred twenty-two amplicons (total of 3 μg DNA) were then pooled in equimolar amounts in a single tube before sequencing by 454 Life Sciences Corp. according to Margulies et al. (9).

Standard bisulfite genomic sequencing analysis. Amplified bisulfite PCR products for *ADAM12* from pooled ALL, FL, and PBL DNA samples described above and an *Sss*I-treated PBL DNA sample were subcloned using the TOPO-TA cloning system (Invitrogen). Plasmid DNA of 30 insert-positive clones for each PCR product was isolated using the Montage Plasmid Miniprep96 kit (Millipore Corp.) and sequenced using an ABI 3730 sequencing system (Applied Biosystems).

Quantitative real-time methylation-specific PCR assay. To further validate the bisulfite sequencing data generated using the 454 GS20 sequencer, DNA was also isolated from 7 normal PBL samples, 19 follicular hyperplasia samples, 29 CLL samples, 13 FL samples, 24 MCL samples, and 10 ALL samples. These DNA samples were treated with sodium bisulfite and examined using quantitative real-time methylation-specific PCR (qMSP) assay for *DLC-1* and *DAPK* methylation. The sequences of qMSP primers and probe sets of *DLC-1* and *DAPK* were reported previously (10, 11). The probes were labeled with two fluorescent dyes (5′-FAM and 3′-BHQ1) and synthesized by Integrated DNA Technologies. In addition, the five bisulfite-treated pooled DNA samples were used for qMSP as described previously (10). The qMSP data specific for methylated DNA were expressed as percent of methylated reference (PMR) values and calculated similarly to a previous report (12). The percentage of fully methylated molecules at a specific locus was calculated by dividing the GENE to *ACTB* ratio of a sample by the GENE to *ACTB* ratio of *Sss*I-treated PBL DNA (methylated reference DNA) and multiplying by 100.

Amplicon identification and dynamic programming alignment programs. Computational analysis of the 454 bisulfite sequencing results for multiple patient groups was conducted using the National Center for Biotechnology Information C toolkit for parsing the files.[4] Each DNA sequence was assigned to 1 of the 25 genes based on the tags and primers. To determine the methylation state for each CpG site in a CGI, the amplicon sequences (i.e., the DNA sequences for the PCR products) for the 25 genes were obtained and then subjected to *in silico* bisulfite conversion. A direct comparison was done between a sequence from the 454 sequencer and its corresponding bisulfite-converted amplicon sequence based on a dynamic programming technique. Detailed procedures and algorithms used for the computational analysis can be found in the Supplementary Materials and Methods.

Cluster analysis. Cluster analysis of the methylation patterns in several amplicons, such as *PON3*, *CYP27B1*, and *LRP1B*, was conducted using Hierarchical Clustering Explorer version 3.5. The clusters were generated by UPGMA (average score clustering), and the similarity between CpG nucleotides on two reads was the count matching of methylated and unmethylated nucleotides or simple matching of nominals.

Statistical analysis. Statistical analyses were carried out using Statistical Analysis System 9.1, Statistical Package for the Social Sciences 14.0, and R 2.4. For the general linear models, some explanatory variables were log transformed to stabilize variance, and residual analysis was used to assess

[4] ftp://ftp.ncbi.nih.gov/toolbox/ncbi_tools

model adequacy. The reported observed significance levels computed in the single nucleotide polymorphism (SNP) analysis were based on asymptotic statistics and are two sided. The expected number of cell counts was always greater than five so that the χ^2 distribution was appropriate.

Results

Preparation of amplicons for 454 sequencing. A massively parallel sequencing-by-synthesis method was used to conduct an ultradeep bisulfite sequencing analysis of 25 gene-related CGIs in five groups of primary cells: (*a*) normal PBL, (*b*) ALL, (*c*) CLL, (*d*) FL, and (*e*) MCL. For each sample group, DNA samples isolated from 10 individuals with the same diagnosis were pooled and bisulfite treated (Fig. 1). The genes included in this study (Supplementary Table S2) were previously shown to be methylated in hematologic neoplasms, including lymphoma and leukemia (10, 13–15). For each gene promoter, a group-specific four-nucleotide tag was added to the 5′ end of each PCR primer (Fig. 1) so that the amplicons from the five groups could be computationally separated after sequencing. Individual amplicons were generated by PCR from each of the five groups for each of the 25 gene-related CGIs. A complete list of the 125 primer pairs is available in Supplementary Table S3, and the locations of the amplicons relative to transcription start sites can be found in Supplementary Fig. S1. Each amplicon was examined by gel electrophoresis, purified, quantified, and then pooled together in equal molar ratios. Three PCRs did not yield high-quality PCR products after several repeated attempts and were excluded from the analysis. The remaining 122 amplicons were sequenced using the GS20 sequencer from 454 Life Sciences.

Mapping the bisulfite sequencing results. All 454 sequences comprised one FASTA file with one sequence read per entry, including quality control information. A total of 294,631 sequences was obtained in a single 5.5-h machine run and included both forward and reverse strand sequences. The average read length was 131 bp (range, 35–300 bp). Bioinformatic analysis consisted of the following three steps: (*a*) match a 454 sequence to a unique primer, (*b*) align the sequence with the *in silico* bisulfite-converted amplicon sequences, and (*c*) compile information from the forward and reverse strands and trim the sequences. A dynamic programming alignment algorithm was used to map each sequence to the *in silico* bisulfite-converted amplicon sequences (see Supplementary Materials and Methods). Of the 294,631 sequences, 288,358 (97.9%) were mapped to a unique amplicon (Table 1). On average, 1,697 sequence reads (including both forward and reverse strands) were obtained for each amplicon. However, some amplicons had many fewer sequence reads than others and occasionally sequences from one strand significantly outnumbered the sequences from the opposite strand. Because the amplicons were examined and quantified before pooling, we suspect that these variations are a result of the series of linker ligation and emulsion PCR amplification steps done in the library preparation protocol used by 454 sequencing. A general linear model was used to determine that the number of sequences obtained is significantly influenced by specific gene, disease type, and the direction of the sequence read. These factors explain 99.6% of the observed variation in the number of reads. The estimated marginal means of the number of reads for each gene by disease type are included in supplementary information (Supplementary Fig. S2). These variations may be caused by the unique sequence structure of each amplicon, including amplicon length, GC content, overall methylation status, and the number of homopolymers present after bisulfite treatment. Because most of them are confounded with gene or gene by disease interaction, they were not included in the model. However, a statistically significant negative correlation was observed between the number of reads and amplicon length ($P < 0.001$) and number of the homopolymer ($P < 0.001$). The sequencing error rate (Table 1) was similar to previously reported values (9). Bisulfite treatment efficiency was determined by calculating the C to T conversion rate for all cytosine bases other than those in CpG dinucleotides (this includes CpA, CpC, or CpT dinucleotides and is from this point referred to as CpH). This was calculated by summing the number of C nucleotides aligned to a CpH and then dividing by the number of C and T nucleotides aligned at CpHs. The primer sequences and filtered reads were not included in this analysis. The bisulfite conversion rate was estimated to be 98.8% (Table 1); however, it could not be

Table 1. Summary of statistics

Statistic	Value
Genes analyzed	25
Patient groups	5
Amplicons generated	122
Mean amplicon length	223.28
Mean number of CpGs per amplicon	17.16
Mean C + G percentage per amplicon	62%
Total sequence reads from 454 sequencer GS20	294,631
Total number of sequences mapped to unique amplicons	288,358
Total number of sequences used for methylation analysis (above 90% sequence identity)	207,011
Total number of sequences used for methylation analysis from forward read	103,755
Total number of sequences used for methylation analysis from reverse read	103,256
Percentage of analyzed CpG sites methylated at or above 0.20 in control group	4.25%
Percentage of analyzed CpG sites methylated at or above 0.20 in noncontrol group	45.64%
Percentage of analyzed CpG sites methylated at or above 0.50 in control group	0.70%
Percentage of analyzed CpG sites methylated at or above 0.50 in noncontrol group	16.24%
CpH conversion efficiency	98.8%
454 sequencing accuracy	99.8%

determined if any of the unconverted cytosines were due to *de novo* CpH methylation.

Quantitative DNA methylation analysis of multiple promoter CGIs. Based on the quality of the alignment, reads with a sequence identity <90% were filtered from the analysis. An example illustrating the alignments of nonfiltered (sequence identity >90%) and filtered (sequence identity <90%) sequences is included in supplementary information (Supplementary Fig. S3). After filtering, a total of 207,011 sequences (70.3%) was used for computing the methylation levels present in the samples. The methylation status of each CpG site in each sequence read was determined based on a C to T conversion at each CpG site on the forward strand and a G to A conversion on the reverse strand. The percentage of methylation at each CpG site within each amplicon, for each sample group, was calculated based on the number of sequences containing methylated CpG sites versus the total number of sequences analyzed (Fig. 2). Most of the CpG sites within the amplicons were analyzed by sequencing from both forward and

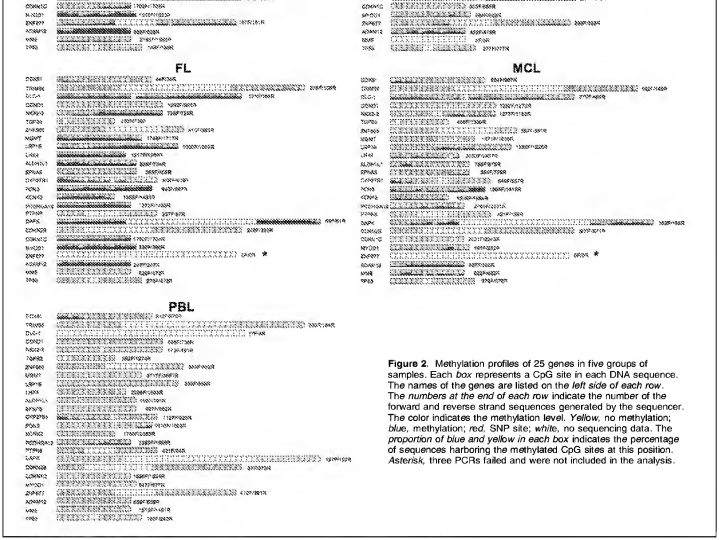

Figure 2. Methylation profiles of 25 genes in five groups of samples. Each *box* represents a CpG site in each DNA sequence. The names of the genes are listed on the *left side of each row*. The *numbers at the end of each row* indicate the number of the forward and reverse strand sequences generated by the sequencer. The color indicates the methylation level. *Yellow*, no methylation; *blue*, methylation; *red*, SNP site; *white*, no sequencing data. The *proportion of blue and yellow in each box* indicates the percentage of sequences harboring the methylated CpG sites at this position. *Asterisk*, three PCRs failed and were not included in the analysis.

reverse directions. However, *TRIM36, ZNF566, PTPN6, DAPK, CDKN2B*, and *ZNF677* had missing data for more than two CpG sites, which is likely a function of the read length limitation of the 454 technology. Overall, the methylation present in the 25 genes analyzed was consistent with what we and others have previously reported (10, 13). However, the results of this study provide greater detail about the quantitative methylation at each single methylcytosine and also provide a methylation profile for each individual fragment analyzed. Although a pooling strategy was used in this pilot study, there was a clear difference between the normal controls and the tumor samples. For instance, the percentage of analyzed CpG sites methylated at or above 20% was 10 times higher in the tumor samples than in the normal PBL samples (45.64% versus 4.25%; see Table 1). Interestingly, the methylation levels at many CpG sites in *ALDH1L1, LRP1B, PON3, PCDHGA12*, and *ADAM12* were exceedingly high (>70%) in ALL and FL samples compared with normal PBL and CLL or MCL samples. This seems to be consistent with our previous findings in which a significantly higher number of methylated genes were identified in FL compared with CLL and MCL (10, 13). Because DNA samples from 10 patients in each diagnostic group were pooled, the presence of CpG methylation in a high proportion of sequence reads for a given amplicon suggests that the majority of patients within the group are methylated at a particular site. Therefore, these unique methylation sites (Fig. 2) have great potential to serve as tumor-specific biomarkers for diagnosis.

To compare the 454 sequencing results with standard bisulfite sequencing, four bisulfite PCR amplicons were generated from an *in vitro*–methylated PBL DNA sample and three pooled DNA samples (PBL, ALL, and FL; see Materials and Methods) analyzed by 454 sequencing using the core primer developed for *ADAM12* but lacking the disease-specific tag designed for the 454 sequencing assay. The four amplicons were cloned and sequenced using the traditional Sanger sequencing method. As shown in Fig. 3, the 454 sequencing results correlate with standard bisulfite sequencing. To examine the quantitative nature of the parallel sequencing method, the 454 sequencing results were compared with qMSP results from *DAPK* and *DLC-1*. The primers and probes used in the qMSP reaction exactly match sequences in the *DAPK* and *DLC-1* amplicons (Supplementary Fig. S4*A* and *C*). The average methylation levels of all CpG sites included in the qMSP primers and probes determined from the 454 sequencing assay were compared with the PMR values obtained from qMSP results and were in good agreement for *DLC-1* but to a lesser degree for *DAPK* (Supplementary Fig. S4*B* and *D*). The 454 sequencing results were also compared with qMSP data from individual patients (Supplementary Fig. S4*B* and *D*). *DLC-1* and *DAPK* methylation were analyzed using qMSP in a total of 102 primary normal and tumor samples (see Supplementary Fig. S5). The average PMR value from each disease group was compared with the qMSP results of pooled DNA samples as well as the sequencing results. A better agreement among three methods was obtained for *DLC-1*. Overall, the confirmatory studies validated the ultradeep bisulfite sequencing results.

DNA methylation patterns in individual CGIs. Twenty of 25 amplicons examined showed an increase in methylation in various types of diseases compared with normal PBL controls. Many also showed quantitatively different levels of methylation between diseases. Particularly, for many of these amplicons, significant increases in methylation densities were observed in FL and ALL compared with CLL and MCL. Because a pooling strategy was used, the contribution of each individual tumor sample to the overall level of methylation could not be assessed. Only 4 of 25 genes (*PON3, CYP27B1, DDX51*, and *PCDHGA12*) had significant methylation (>20%) at certain CpG sites in the normal controls but in each of these genes there was an increase in the levels of methylation present in the tumor samples. Interestingly, a demarcation along the border of the CGIs of *PON3* and *CYP27B1* was observed in control normal PBLs, with little or no methylation in the middle of the CGI but low to moderate levels (20–44%) of methylation along the borders. Similarly, this demarcation line was clearly seen in some tumor samples, but a significant increase in methylation density at the border of the CGI was observed (Fig. 2). Each *PON3* and *CYP27B1* sequence obtained was analyzed using a

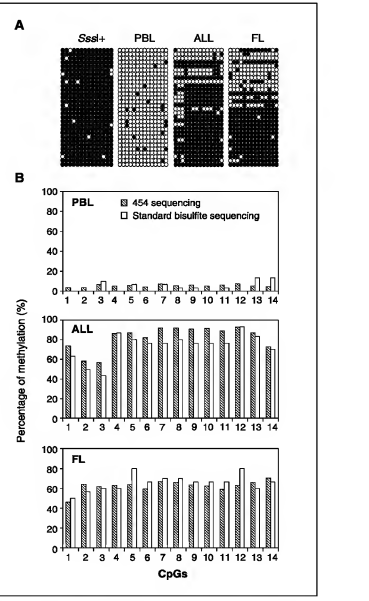

Figure 3. Validation of the massively parallel bisulfite sequencing results of *ADAM12* by standard bisulfite sequencing. *A*, standard bisulfite sequencing of *ADAM12* in the pooled DNA (ALL, FL, and PBL) samples studied by 454 sequencing and an *SssI*-treated PBL DNA sample. *Row*, sequence of individual clones. ●, methylated CpG sites; ○, unmethylated CpG sites. *B*, comparison of 454 sequencing results with standard bisulfite sequencing.

clustering program (Fig. 4). Compared with control normal PBL samples, the methylation patterns for *PON3* in FL and ALL seem to suggest a progressive spreading of methylation from the border of the CGI toward the center of the island (Fig. 4*A*). In these cases, the demarcation line became less distinct and showed high levels (>80%) of methylation along the border accompanied by an increase in the number of methylated CpG sites in the promoter area. Similar gradual changes in methylation were observed for *CYP27B1* in FL and ALL (Fig. 4*B*). These results seem to support the theory that methylation spreads from the outside of the CGI toward the center of the island.

Analysis of SNPs in amplicons. 454 sequencing has been used for deep sequencing and for identifying rare mutations (16). Because bisulfite treatment only modifies unmethylated cytosine, but not adenine, guanine, or thymine, genetic changes in sequences other than CpG dinucleotides were analyzed. A search of the SNP database available at the University of California at Santa Cruz genome Web site identified 11 SNPs in 8 of the 25 genes analyzed in our study. One of these (rs4646696) was present within a CpG site in the *ALDH1L1* amplicon (Fig. 2). No disease-specific association between any of the published SNPs and lymphoma or leukemia was identified. However, one G→C polymorphism (rs1375610) created an additional CpG site in the *LRP1B* amplicon. A clustering algorithm was used to analyze the methylation patterns of sequences with either the C allele or the G allele. As shown in Fig. 5*A*, the C allele sequences can clearly be separated by the methylation status of the SNP position. An overwhelming majority of the fragments with a methylated cytosine at the SNP site were also methylated at most of the remaining CpG sites within the amplicon. To quantify the association between the G→C polymorphism and methylation status within *LRP1B*, odds ratios were computed for each CpG site within the amplicon. A positive log odds ratio indicates that methylation, at that position, is associated with the G→C polymorphism. As shown in Fig. 5*B*, methylation was virtually always associated with the SNP (15 of 16) and statistical significance was found at the vast majority of positions (13 of 16). Averaging across all positions, the odds of methylation were more than twice as high for the C allele versus the G allele. Despite the fact that the statistical significance in the case of pooled samples with many repeated measurements is an optimistic estimate that should be interpreted loosely, the potential effect of this SNP on the overall methylation of *LRP1B* warrants further investigation.

Discussion

We have shown in this report that a massively parallel pyrosequencing (454 sequencing) technology can be used in the high-throughput sequencing of bisulfite PCR amplicons. Currently, several other parallel sequencing technologies are available or under development from Solexa, Agencourt, and Helicos. At present, the 454 platform generates the longest sequence reads (average 100 bp compared with 25–35 bp with other technologies) and therefore was the best suited for this experimental design. However, the concept of this approach can be applied to other emerging sequencing platforms if read lengths are significantly improved. In this pilot study, a pooling strategy combined with a single library preparation was used to sequence multiple disease types and individuals in a single run. Because the GS20 sequencer is capable of running 1, 4, or 16 different samples in a single machine run, a combinatory experimental design using pooling and primer tagging strategies could potentially maximize the number of genes or samples analyzed in a single sequencing run. This would also confer a considerable reduction in cost ($0.06 versus $1–5 for standard bisulfite sequencing) and in labor.

Bisulfite-based approaches are dependent on complete chemical conversion of unmethylated cytosines in the DNA. Using an ultradeep bisulfite sequencing approach, the conversion of a large

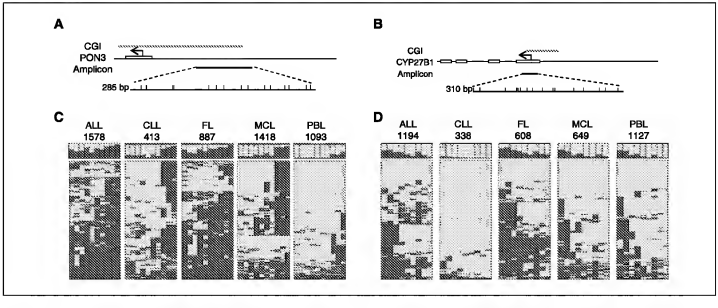

Figure 4. The spreading of DNA methylation from the periphery toward the center of select CGIs. *A* and *B*, *top*, location of the amplicons studied relative to the transcription start site and within the CGI. The *bar underneath each amplicon* illustrates the relative location of each CpG within the amplicon (*vertical bars*). *C* and *D*, cluster analysis of the bisulfite sequencing data of the *PON3* and *CYP27B1* genes. Disease types are labeled on the *top* of each panel. The number of sequence reads used for cluster analyses is listed under the disease label. The color indicates methylation status of each CpG site. *Blue*, methylation; *yellow*, no methylation. Each *column* represents a CpG site within each amplicon. The *bar graph on top* shows the overall methylation level of each CpG site for the five pooled DNA samples. The heat maps on the *bottom* represent cluster results.

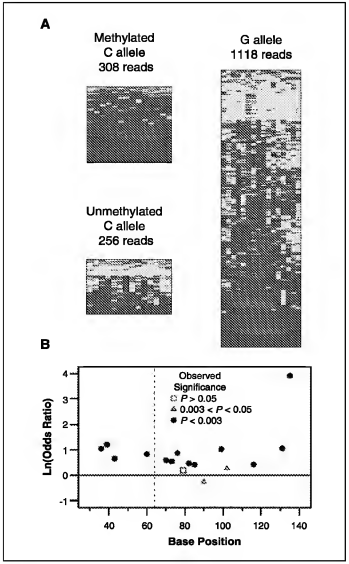

Figure 5. Cluster analysis of the bisulfite sequencing data of the *LRP1B* gene in ALL samples. The color indicates methylation status of each CpG site. *Blue*, methylation; *yellow*, no methylation; *white*, no methylation data. Each *column* represents a CpG site within the *LRP1B* amplicon. Each *row* represents a sequencing read. *A*, *top*, the 1,682 sequence reads were first grouped based on the SNP genotype and the methylation status of the SNP site and then clustered using the same clustering algorithms. *B*, at each relevant base position in the *LRP1B* amplicon sequence, the natural logarithm of the odds ratio (odds of methylation with C allele/odds of methylation with G allele) is given. *Points lying above the solid horizontal line*, positive association between the G→C polymorphism and methylation; *points below the line*, negative association; *broken vertical line*, location of the polymorphism. *Points represented by a solid circle* have a *P* value of <0.003. These points are significant with a family-wise error rate of 5% using Bonferroni's correction.

number of cytosine bases was assessed providing a reliable estimation of the bisulfite conversion rate. Using the current protocol, a 98.8% bisulfite conversion efficiency was achieved (Table 1). Bisulfite treatment reduces the sequence diversity and generates many AT-rich sequences in addition to creating homopolymer stretches that can influence the accuracy of the sequences generated using the 454 technology. However, our results corroborate claims made by the company that this system is capable of generating accurate sequences even in AT-rich regions of the genome (9). Furthermore, the dynamic programming algorithms used in this study were able to accurately align fragments even if there were gaps present in a homopolymeric region. Despite the many successes encountered, some amplicons did not produce a sufficient number of high-quality sequence reads for the quantification of methylcytosine methylation (i.e., *DLC-1* in PBL). This is likely due to the amplification bias at the library preparation step used at 454, resulting in a bias of the fragments generated during the emulsion PCR reaction. Although this is a major concern, only 2 of 122 amplicons had such a problem in this pilot study.

The ultradeep bisulfite sequencing, to our knowledge, generated the most comprehensive quantitative analysis of DNA methylation patterns in multiple gene promoters and multiple tumor types. The results show that the distribution of methylated sites is not equally represented within a CGI. For example, progressive spreading of methylation was identified from the boundary of the CGI and gradually moving toward the transcription start site of genes, such as *PON3* and *CYP27B1*. In the control normal PBLs analyzed, low levels of methylation were seen in the 3′ end (first exon or intron) or 5′ end (upstream) boundary of the CGI but not in the promoter area. A methylation wave may occur, progressively extending from the boundary toward the promoter area of some CGIs in lymphomas and leukemias. These results further confirm the previous observations of the progressive spreading of *RASSF1A* methylation in breast cancer *in vivo* (17) and of the spreading of methylation in the *E-cad* and *VHL* CGIs in cultured fibroblasts overexpressing DNA methyltransferase *DNMT1* (18). Additionally, the spreading of methylation seems to correlate with the overall increase in promoter methylation in specific disease types. For instance, the spreading of methylation in ALL and FL is more significant than MCL and CLL and this is correlated with higher numbers of methylated genes in ALL and FL. The underlying mechanism of this epigenetic event is not clear but could relate to lymphoma or leukemia genesis. One common feature of ALL and FL is that, although these cells arise from two distinct maturation stages of B-cell development, the cells from both tissues are undergoing rapid mutations and rearrangements in DNA under normal conditions. Because DNA methyltransferases are associated with DNA repair (19), it is possible that DNA methyltransferases are actively expressed during these two stages playing a mechanistic role in the overabundance of methylated genes present in these two diseases. Further, *DNMT1* expression is significantly higher in normal germinal center B cells than in normal naive, memory, mantle zone, or marginal zone B cells (20). Consequently, aberrantly functioning DNMT1 could explain the overabundance of methylated genes in FL (derived from germinal center B cells) when compared with CLL and MCL (derived from marginal zone B cells).

Because of limited sample availability, gene expression analyses could not be conducted using matched RNA samples. However, the expression levels of *DLC-1*, *LRP1B*, *CYP27B1*, *KCNK2*, *PCDHGA12*, *DDX51*, *CCND1*, *p57*, and *MME* (*CD10*) in lymphoma and leukemia cell lines were previously characterized and it was shown that *in vitro* treatment with a demethylating agent could reactivate these genes in cell lines possessing hypermethylated gene promoters (10, 13, 21, 22). Furthermore, the expression levels of *DLC-1* and *LRP1B* were previously assessed in primary lymphoma samples and a reciprocal correlation was found between hypermethylation and gene expression, indicating that promoter hypermethylation may play a role in the down-regulation of *DLC-1* and *LRP1B* gene expression in primary lymphoma samples (10, 13).

With recent developments in epigenomic technologies, our group and others have discovered many novel methylated promoter CGIs that may be tumor specific and have great potential as epigenetic biomarkers. qMSP is a promising methylation analytic method that could be used for developing clinical diagnostic assays but is limited to the analysis of only a few CpG sites within a given CGI. Therefore, it is critical that the CpG sites selected for designing such assays are not biased. In this study, we generated a comprehensive, in-depth analysis of the promoter methylation identified by our microarray studies at a single molecule level and a single methylcytosine resolution. The information obtained can be used to guide the design of qMSP assays. In fact, several genes studied, such as *ADAM12*, *ALDH1L1*, and *LRP1B*, may be excellent candidates for developing epigenetic biomarkers, which show a far greater level of methylation than several selected candidate genes that were previously shown to be methylated in lymphomas and leukemias, such as *p15* (*CDKN2B*) and *SHP-1* (*PTPN6*). In addition, this ultradeep sequencing approach is capable of identifying genetic mutations and providing genotype information that could potentially link genetic and epigenetic data, leading to the development of comprehensive markers for disease classification and diagnosis.

Although a great potential for the application of next-generation sequencing technologies in epigenetic studies was shown in this study, several issues need to be addressed to fully use the throughput capabilities of high-throughput parallel sequencing instruments. In this proof-of-concept study, a large number of individual PCR assays were done before pooling for sequencing. Alternately, one could incorporate the Gene-Collector method, which generates uniformly distributed multiplexed amplification products with less bias (23), to increase the efficiency of amplicon generation. Furthermore, careful optimization of the amplification protocols, designing PCR amplicons of consistent length, and addition of spiked controls should be considered in future studies.

In summary, this study provides evidence that high-throughput parallel bisulfite sequencing can survey DNA methylation in genomic regions of interest in an ultradeep fashion and at a single methylcytosine resolution. Such a thorough analysis is expected to provide insights into the progressive nature of aberrant DNA methylation and its relationship to transcriptional silencing in the neoplastic process and to assist in the design of quantitative epigenetic biomarkers that can be used in diagnostic assays.

Acknowledgments

Received 3/19/2007; revised 6/15/2007; accepted 7/3/2007.

Grant support: National Cancer Institute grants CA123018 (H. Shi) and CA100055 and CA097980 (C.W. Caldwell), NIH Biomedical and Health Informatics Research Training Program LM07089 (R.S. Kramer), and National Science Foundation grant ITR-IIS-0407204 (D. Xu). C.W. Caldwell is Cancer Research Center Missouri Chair in Cancer Research.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

We thank Dr. Michael X. Wang for kindly providing several DNA samples used in this study.

References

1. Bird A. The essentials of DNA methylation. Cell 1992;70:5–8.
2. Robertson KD, Jones PA. DNA methylation: past, present and future directions. Carcinogenesis 2000;21:461–7.
3. Craig JM, Bickmore WA. The distribution of CpG islands in mammalian chromosomes. Nat Genet 1994;7:376–82.
4. Jones PA, Baylin SB. The fundamental role of epigenetic events in cancer. Nat Rev Genet 2002;3:415–28.
5. Grunau C, Clark SJ, Rosenthal A. Bisulfite genomic sequencing: systematic investigation of critical experimental parameters. Nucleic Acids Res 2001;29:E65.
6. Sandovici I, Leppert M, Hawk PR, Suarez A, Linares Y, Sapienza C. Familial aggregation of abnormal methylation of parental alleles at the IGF2/H19 and IGF2R differentially methylated regions. Hum Mol Genet 2003;12:1369–78.
7. Eckhardt F, Lewin J, Cortese R, et al. DNA methylation profiling of human chromosomes 6, 20 and 22. Nat Genet 2006;38:1378–85.
8. Rakyan VK, Hildmann T, Novik KL, et al. DNA methylation profiling of the human major histocompatibility complex: a pilot study for the human epigenome project. PLoS Biol 2004;2:e405.
9. Margulies M, Egholm M, Altman WE, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature 2005;437:376–80.
10. Shi H, Guo J, Duff DJ, et al. Discovery of novel epigenetic markers in non-Hodgkin's lymphoma. Carcinogenesis 2007;28:60–70.
11. Yegnasubramanian S, Kowalski J, Gonzalgo ML, et al. Hypermethylation of CpG islands in primary and metastatic human prostate cancer. Cancer Res 2004;64:1975–86.
12. Widschwendter M, Siegmund KD, Muller HM, et al. Association of breast cancer DNA methylation profiles with hormone receptor status and response to tamoxifen. Cancer Res 2004;64:3807–13.
13. Rahmatpanah FB, Carstens S, Guo J, et al. Differential DNA methylation patterns of small B-cell lymphoma subclasses with different clinical behavior. Leukemia 2006;20:1855–62.
14. Esteller M. Profiling aberrant DNA methylation in hematologic neoplasms: a view from the tip of the iceberg. Clin Immunol 2003;109:80–8.
15. Taylor KH, Pena-Hernandez KE, Davis JW, et al. Large-scale CpG methylation analysis identifies novel candidate genes and reveals methylation hotspots in acute lymphoblastic leukemia. Cancer Res 2007;67:2617–25.
16. Thomas RK, Nickerson E, Simons JF, et al. Sensitive mutation detection in heterogeneous cancer specimens by massively parallel picoliter reactor sequencing. Nat Med 2006;12:852–5.
17. Yan PS, Shi H, Rahmatpanah F, et al. Differential distribution of DNA methylation within the RASSF1A CpG island in breast cancer. Cancer Res 2003;63:6178–86.
18. Graff JR, Herman JG, Myohanen S, Baylin SB, Vertino PM. Mapping patterns of CpG island methylation in normal and neoplastic cells implicates both upstream and downstream regions in *de novo* methylation. J Biol Chem 1997;272:22822–9.
19. Li YQ, Zhou PZ, Zheng XD, Walsh CP, Xu GL. Association of Dnmt3a and thymine DNA glycosylase links DNA methylation with base-excision repair. Nucleic Acids Res 2007;35:390–400.
20. Martin-Subero JI, Ballestar E, Esteller M, Siebert R. Towards defining the lymphoma methylome. Leukemia 2006;20:1658–60.
21. Guo J, Burger M, Nimmrich I, et al. Differential DNA methylation of gene promoters in small B-cell lymphomas. Am J Clin Pathol 2005;124:430–9.
22. Taylor KH, Liu J, Guo J, Davis JW, Shi H, Caldwell CW. Promoter DNA methylation of CD10 in lymphoid malignancies. Leukemia 2006;20:1910–2.
23. Fredriksson S, Baner J, Dahl F, et al. Multiplex amplification of all coding sequences within 10 cancer genes by Gene-Collector. Nucleic Acids Res 2007;35:e47.